THE SENSITIVITY OF TWENTY MEASURES OF PILOT MENTAL WORKLOAD IN A SIMULATED ILS TASK

by

Walter W. Wierwille and Sidney A. Connor
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

Twenty workload estimation techniques were compared in terms of their sensitivity to changes in pilot loading in an ILS task. The techniques included opinion measures, spare mental capacity measures, physiological measures, eye behavior measures, and primary task measures. Loading was treated as an independent variable and had three levels: low, medium, and high. The load levels were obtained by a combined manipulation of windgust disturbance level and simulated aircraft pitch stability. Six instrumented-rated pilots flew a moving-base general aviation simulator in four sessions lasting approximately three hours each. Measures were taken between the outer and middle markers.

Two opinion measures, one spare mental capacity measure, one physiological measure, and one primary task measure demonstrated sensitivity to loading in this experiment. These measures were: Cooper-Harper ratings, WCI/TE ratings, time estimation standard deviation, pulse rate mean, and control movements per unit time. The Cooper-Harper ratings, WCI/TE ratings, and control movements demonstrated sensitivity to all levels of load, whereas the time estimation measure and pulse rate mean showed sensitivity to some load levels.

The results of this experiment demonstrate that sensitivities of workload estimation techniques vary widely, and that only a few techniques appear to be sensitive in this type of ILS task, which emphasizes psychomotor behavior.

INTRODUCTION

One of the major problems in mental workload estimation is the lack of available information on the sensitivity of various workload estimation techniques [1,2]. When a researcher or human factors engineer needs to assess workload in a given experimental situation, it is not clear which technique or techniques should be used [3]. The danger is that insensitive techniques may be used. If so, experimental results will show no differences in workload when in fact there are differences.

Sensitivity in regard to workload estimation can be defined as the relative ability of a given workload estimation technique to discriminate statistically significant differences in operator loading. High sensitivity requires

discriminable changes in the score means as a function of load level and low variation of the scores about the means. When sensitivity is defined in this way, it becomes subject to experimental determination. Based on experiments that emphasize specific operator behaviors, it should be possible to predict which given techniques are sensitive.

An experiment directed at evaluating the sensitivity of workload estimation techniques in a psychomotor task has been completed and is reported briefly in this paper. An ILS piloting task was used for the evaluation. (For a more detailed discription of the experiment and results, see reference [4]).

EXPERIMENT

Subjects

Six male instrument-rated pilots served as subjects in this experiment. The flight time of the subjects ranged from 500 to 2700 hours with a mean of 1300 hours.

Apparatus

The primary apparatus in this experiment was a modified flight task simulator (Singer Link, Inc., General Aviation Trainer, GAT-1B). The simulator had three degrees of freedom of motion (roll, pitch, and yaw). Transulucent blinders were used to cover the windows of the simulator to reduce outside distractions and cues and to aid in the control of cockpit illumination.

Several modifications to the flight simulator were made for the experiment. These modifications permitted primary task load manipulation, secondary task operations, response measurement, and scoring. Primary task load manipulation was accomplished by changing aircraft pitch stability and random wind-gust disturbance level simultaneously. Three load conditions were developed: low, medium, and high, as shown in Table 1. Table 2 provides a list of the workload measurement techniques selected for inclusion in the present study.

Experimental Design

A complete 3 x 20 within-subject design was used for the sensitivity analysis. Load was the factor with three levels. Measurement technique (Table 2) was the factor with twenty levels.

Workload measures from different techniques were taken simultaneously on some of the data collection runs. Only those measures which were not likely to affect each other were taken simultaneously. Table 3 shows the scheme used for combining different measurement techniques for data collection. The combination of measurement techniques shown in the table was, to an extent, based on previous investigations of workload. Hicks and Wierwille's [3] study supported the combination in condition 2. The two rating scales were administered

in separate measurement conditions to prevent the ratings on one scale from biasing the ratings on the other scale. The secondary task measures were divided among several conditions because of potential intrusion and interference. Vocal measures were recorded from the two secondary tasks which required a verbal response as per Schiflett and Loikith's [5] recommendation.

It should be noted that primary task measures were recorded on all subjects and on all data collection flights for the intrusion analysis. However, only data from measurement condition 1 were used for the sensitivity analysis of the primary task measures.

General Procedure

After receiving instructions, subjects flew nine familiarization flights in the simulator. These flights were similar, but not the same as, the data collection flights. All subjects flew the familiarization flights in the same order. Steady crosswinds were introduced for each run, and subjects were given heading corrections.

After the familiarization session, the subjects participated in three data collection sessions. The familiarization session and each data collection session were held on a different day.

Each data collection session consisted of two sets of a warm-up practice flight and three data collection flights. The practice flight was the same as the first data collection flight. Since the data collection flights were counterbalanced, equal amounts of practice were provided for the low, medium, and high load conditions. The data collection flights also contained steady crosswind conditions, for which the subject was given heading corrections. The purpose of introducing steady crosswinds was to disguise the load conditions, thereby requiring subjects to fly each flight as a separate entity.

Flight Task Procedures

The flight task in this experiment was an ILS approach in the Singer Link GAT-1B aircraft simulator. Prior to the beginning of a flight, the simulated aircraft was positioned on the ground 5 miles outbound from the outer marker on the 108 degree radial, heading into the wind. When ready to begin, the experimenter informed the subject of the wind direction and speed, and gave him a heading correction for the crosswind. When contacted by the experimenter, the subject took off and climbed to 2000 feet. The subject then flew directly to the outer marker by following the localizer at 100 miles per hour until the glide slope was intercepted. Upon interception of the glide slope, the subject reduced airspeed to 80 miles per hour and proceeded down the glide slope while following the localizer to a landing. Data were recorded between the outer and middle markers. For the opinion measures, subjects gave ratings for the flight segment between the outer and middle markers immediately after landing and parking the simulated aircraft.

RESULTS

The computed scores for each technique were first converted to Z-scores (normalized scores) so that technique measure units would not affect the sensitivity analysis. Subsequently, an overall analysis of variance was performed on the scores. Since Z-scores were used, a technique main effect was not possible. A significant main effect of load was found, F (2,10) = 5.34, $p < 0.0001$, and a significant load by technique interaction was found, F (38,190) = 2.76, $p \leq 0.05$.

The load by technique interaction indicated that the measurement techniques were differentially sensitive to load. Therefore, individual ANOVAs were used to isolate the sensitive techniques.

The individual ANOVAs indicated that five of the twenty measures were sensitive. They were the Cooper-Harper scale F (2,10) = 16.39, $p = 0.0007$; the Workload-Compensation-Interference/Technical Effectiveness (SCI/TE) scale, F (2,10) = 31.15, $p < 0.0001$; the time estimation standard deviation, F (2,10) = 5.69, $p = 0.022$; the pulse rate mean, F (2,10) = 8.89, $p = 0.006$; and the control movements measure, F (2,10) = 33.34 $p < 0.0001$. The normalized means for each technique are plotted in Figures 1 through 5 as a function of load.

Newman-Keuls comparisons were then performed on the normalized means of the sensitive measures. The comparisions included low vs. medium, medium vs. high, and low vs. high load conditions. Results indicated that all differences were significant at $p < 0.05$, except for pulse-rate mean (low vs. medium and medium vs. high) and time estimation standard deviation (low vs. high).

A logical classification of techniques based on demonstrated sensitivity was generated from an examination of the Newman-Keuls comparisons, as shown in Table 4. Techniques which demonstrated sensitivity to all pairs of load conditions (i.e., low vs. medium, medium vs. high, and low vs. high) were included in class I. These measures are preferred over other techniques which demonstrated only partial sensitivity, or no sensitivity in the present study. Techniques which showed sensitivity to some differences in load conditions (but not all) were included in class II. These measures are less preferred than class I techniques, but are more preferred than class III techniques. Class III techniques did not demonstrate sensitivity to load in the present study. This class includes all techniques except those in class I and class II.

One possible reason that only five of the twenty workload assessment techniques demonstrated sensitivity in the present study is that the other techniques simply required a greater number of subjects to show a significant effect of load. It is possible to estimate the sample size required to detect a reliable load effect for a given workload assessment technique at specified levels of significance and power. These calculations were performed for techniques which did not demonstrate sensitivity in the present study, to provide an indication of the practical costs of achieving statistical significance. The procedure used for estimating the sample size required for finding sensi-

tivity is described by Bowker and Lieberman [6]. Sample sizes were estimated for a significance level of 0.05 and for a power of approximately 0.80. The results of these estimates are presented in Table 5.

CONCLUSIONS

This study has shown that five measures of workload estimation were sensitive indicators of load in a piloting task that is predominantly psychomotor in nature. Another fifteen measures, believed to be "good" measures of workload, showed no reliable effect. The main conclusion that must be drawn from the study is that few measures are sensitive to psychomotor load.

Of the five techniques demonstrating sensitivity, only three exhibited monotonic score increases with load as well as statistically reliable differences between all pairs of load levels. Consequently, only the three meet all criteria for sensitivity to psychomotor load. These class I techniques are the ones that are recommended for measurement of psychomotor load:

Cooper/Harper ratings,
WCI/TE ratings, and
Control movements per second.

The other two techniques showed sensitivity to psychomotor load, but did not discriminate between all pairs of load levels. These class II techniques are:

Time estimation standard deviation, and
Pulse rate mean.

These measures would be helpful in evaluating psychomotor load, but they should not be relied on exclusively. At least one class I technique should also be used in conjunction with these measures.

It is worth noting that only two opinion measures were taken in the present experiment, and both proved sensitive. This suggests that well-designed rating scales are among the best of techniques for evaluating psychomotor load. In regard to the primary task measures, the control movements measure alone was sensitive. However, this measure is also the only primary task measure which reflected "strategy" of the pilot. Consequently, one could speculate that selecting a primary task measure that reflects strategy will most likely result in good sensitivity.

Fifteen (techniques) measures showed no reliable change as a function of load. When these fifteen measures were subjected to a power analysis to determine sample size, the number of subjects required ranged from 12 to well over 100 (Table 5). One can only conclude that at best the fifteen measures, as taken, are much less sensitive to psychomotor load than the five appearing in Classes I and II. Of course, there is always the possibility that the measures would be sensitivite to loading along other dimensions of human performance, such as psychomotor tasks of a different nature, or mediational or cognitive tasks, for example.

In general, the results of the experiment show that there are wide variations in the sensivity of workload estimation measures. Great care must be taken in selecting measures for a given experiment. Otherwise, it is possible that no changes in workload will be found, when indeed there are changes.

REFERENCES

1. Wierwille, W. W. and Williges, R. C. Survey and analysis of operator workload assessment techniques. Blacksburg, Virginia: Systemetrics, Inc. Report No. S-78-101, September, 1978.

2. Wierwille, W. W. and Williges, B. H. An annotated bibliography on operator mental workload assessment. Patuxent River, Maryland: Naval Air Test Center Report No. SY-27R-80, March, 1980.

3. Hicks, T. G. and Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving base driving simulator. Human Factors, 1979, 21, 129-143.

4. Connor, S. A. and Wierwille, W. W. Comparative evaluation of twenty pilot workload assessment measures using a psychomotor task in a moving base simulator. Moffett Field, CA: NASA-Ames Research Center, (Forthcoming report).

5. Schiflett, S. G. and Loikith, G. J. Voice stress as a measure of operator workload. Patuxent River, Maryland: Naval Air Test Center, Technical Memorandom TM 79-3 SY, December 31, 1979.

6. Bowker, A. H. and Lieberman, G. J. Engineering statistics. New Jersey: Prentice-Hall, Inc., 1959.

ACKNOWLEDGEMENTS

The authors wish to thank Mrs. Sandra Hart, NASA-Ames Research Center, for helpful technical suggestions. This work was sponsored under NASA grant NAG2-17.

TABLE 1

Primary Task Load Conditions

	LOAD CONDITION		
	Low	Medium	High
RANDOM GUST LEVEL	Low	Medium	High
Estimated Std. Dev. (mph)	0	2.7	5.9
PITCH STABILITY	High	Medium	Low
a. Control input to pitch rate output equivalent gain (degrees/s per % of control range)	0.522	3.560	7.83
b. Control input to pitch rate output equivalent time constant(s)	0.097	0.660	1.45

TABLE 2

Workload Assessment Techniques Which Were Tested in the Present Experiment

OPINION	
1.	Cooper-Harper Scale
2.	WCI/TE Scale
SPARE MENTAL CAPACITY	
3.	Digit Shadowing (% errors)
4.	Memory Scanning (Mean time)
5.	Mental Arithmetic (% errors)
6.	Time Estimation Mean (Seconds)
7.	Time Estimation Standard Deviation (Seconds)
8.	Time Estimation Absolute Error (Seconds)
9.	Time Estimation RMS error (Seconds)
PHYSIOLOGICAL	
10.	Pulse Rate Mean (Pulses per minute)
11.	Pulse Rate Variability (Pulses per minute)
12.	Respiration Rate (Breath cycles per minute)
13.	Pupil Diameter (Normalized units)
14.	Voice Pattern (Digit Shadowing Task)
15.	Voice Pattern (Mental Arithmetic Task)
EYE BEHAVIOR	
16.	Eye Transition Frequency (Transitions per minute)
17.	Eye Blink Frequency (Blinks per minute)
PRIMARY TASK	
18.	Localizer RMS Angular Position Error (Degrees)
19.	Glide Slope RMS Angular Position Error (Degrees)
20.	Control Movements per second (Aileron + Elevator + Rudder)

TABLE 3

Combination of Measurement Techniques
for Data Collection

Measurement Condition	Measurement Techniques
1.	Cooper-Harper Scale Pupil Diameter Eye Transition Frequency Eye Blink Frequency Localizer RMS Error Glide Slope RMS Error Control Movements
2.	WCI/TE Scale Pulse Rate Mean Pulse Rate Variability Respiration Rate
3.	Digit Shadowing Voice Pattern
4.	Memory Scanning
5.	Mental Arithmetic Voice Pattern
6.	Time Estimation (Mean) (Std. Dev.) (Abs. Error) (RMS Error)

TABLE 4

Logical Classification of Techniques
Based on Demonstrated Sensitivity

Class I: Complete Sensitivity Demonstrated
- Cooper-Harper Scale
- WCI/TE Scale
- Control Movements/Unit Time

Class II: Some Sensitivity Demonstrated
- Time Estimation Standard Deviation*
- Pulse Rate Mean**

Class III: Sensitivity Not Demonstrated
- All Other Techniques (See Table 5)

*Double valued function
**Limited sensitivity

TABLE 5

Estimated Sample Sizes Required for Achieving a Significant Load Effect for Techniques not Demonstrating Sensitivity

Technique	Estimated Sample Size
SPARE MENTAL CAPACITY	
Digit Shadowing	18
Memory Scanning	>100
Mental Arithmetic	25
Time Estimation (Mean)	53
Time Estimation (Abs. Error)	>100
Time Estimation (RMS Error)	53
PHYSIOLOGICAL	
Pulse Rate Variability	45
Respiration Rate	15
Pupil Diameter	>100
Speech Pattern (D. Shadow.)	28
Speech Pattern (M. Arith.)	>100
EYE BEHAVIOR	
Eye Transition Frequency	42
Eye Blink Frequency	25
PRIMARY TASK	
Localizer RMS Error	12
Glide Slope RMS Error	41

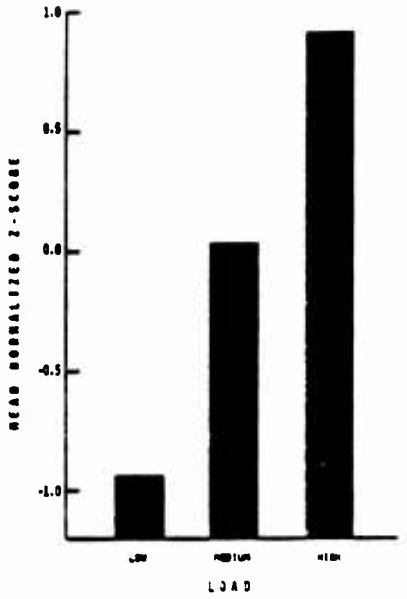

Figure 1. Mean normalized scores for the Cooper-Harper rating scale measure plotted as a function of load.

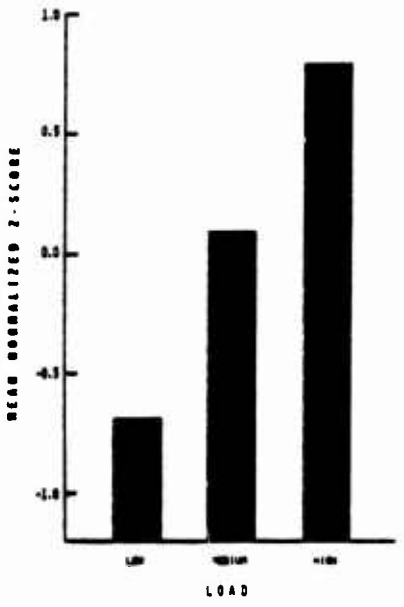

Figure 2. Mean normalized scores for the WCI/TE rating scale measure plotted as a function of load.

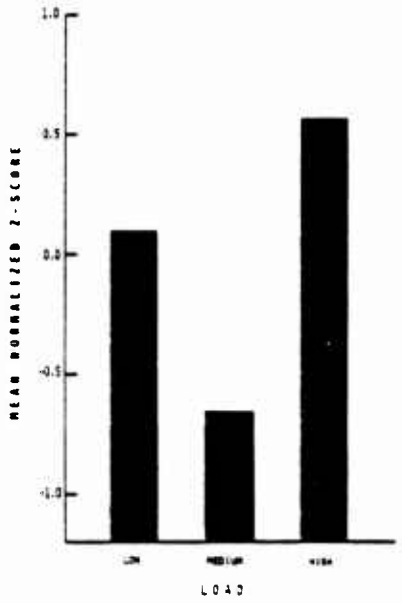

Figure 3. Mean normalized scores for the time estimation standard deviation measure plotted as a function of load.

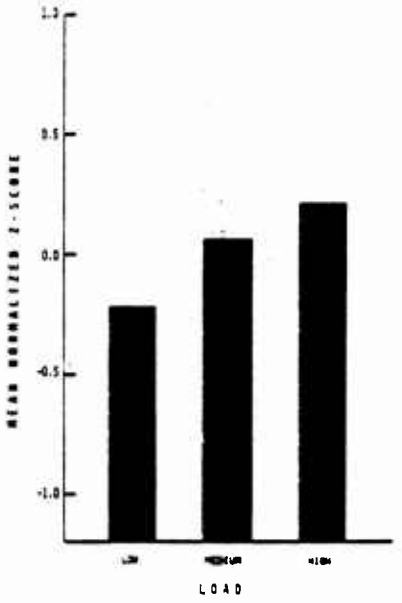

Figure 4. Mean normalized scores for the pulse rate mean measure plotted as a function of load.

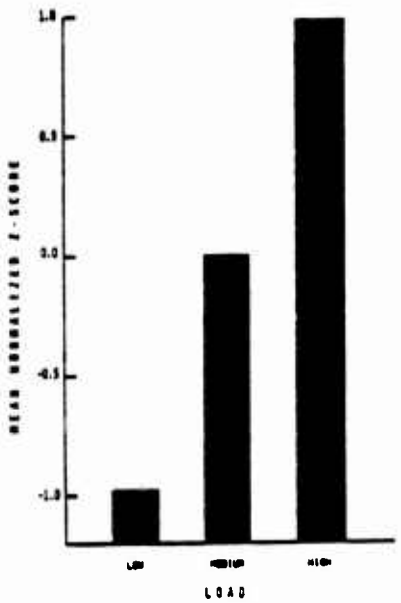

Figure 5. Mean normalized scores for the control movements measure plotted as a function of load.